Statistics for Social and Behavioral Sciences

Michael J. Kolen
Robert L. Brennan

Test Equating, Scaling, and Linking

Methods and Practices

Third Edition

Springer

Statistics for Social and Behavioral Sciences

Series editors

S. E. Fienberg
W. J. van der Linden

For further volumes:
http://www.springer.com/series/3463

Michael J. Kolen · Robert L. Brennan

Test Equating, Scaling, and Linking

Methods and Practices

Third Edition

Springer

Michael J. Kolen
Iowa Testing Programs
University of Iowa
Iowa City, IA
USA

Robert L. Brennan
CASMA
University of Iowa
Iowa City, IA
USA

ISBN 978-1-4939-0316-0 ISBN 978-1-4939-0317-7 (eBook)
DOI 10.1007/978-1-4939-0317-7
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013956631

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Amy, Raychel, and Daniel

—M. J. K.

To Cicely and Sean

—R. L. B.

Preface

Prior to 1980, the subject of equating was ignored by most people in the measurement community except for psychometricians, who had responsibility for equating. Beginning in the early 1980s, the importance of equating was recognized by a broader spectrum of people associated with testing. This increased attention to equating is attributable to at least three developments. First, there continues to be an increase in the number and variety of testing programs that use multiple forms of tests, and the testing professionals responsible for such programs have recognized that scores on multiple forms should be equated. Second, test developers and publishers often have referenced the role of equating in arriving at reported scores to address a number of issues raised by testing critics. Third, the accountability movement in education and issues of fairness in testing have become much more visible. These developments have given equating an increased emphasis among measurement professionals and test users.

In addition to statistical procedures, successful equating involves many aspects of testing, including procedures to develop tests, to administer and score tests, and to interpret scores earned on tests. Of course, psychometricians who conduct equating need to become knowledgeable about all aspects of equating. The prominence of equating, along with its interdependence with so many aspects of the testing process, also suggests that test developers and all other testing professionals should be familiar with the concepts, statistical procedures, and practical issues associated with equating.

Before we published the first edition in 1995, the need for a book on equating became evident to us from our experiences in equating hundreds of test forms in many testing programs, in training psychometricians to conduct equating, in conducting seminars and courses on equating, and in publishing on equating and other areas of psychometrics. Our experience suggested that relatively few measurement professionals had sufficient knowledge to conduct equating. Also, many did not fully appreciate the practical consequences of various changes in testing procedures on equating, such as the consequences of many test-legislation initiatives, the use of constructed-response items in assessments, and the introduction of computer-based test administration. Consequently, we believed that measurement professionals needed to be educated in equating methods and practices; the 1995 book was intended to help fulfill this need. Although several general published references on equating existed at the time (e.g., Angoff 1971;

Harris and Crouse 1993; Holland and Rubin 1982; Petersen et al. 1989), none of them provided the broad, integrated, in-depth, and up-to-date coverage of the first edition of this book.

After the publication of the first edition in 1995, a large body of new research was published. Much of this work was in technical areas that include smoothing in equipercentile equating, estimation of standard errors of equating, and the use of polytomous item response theory methods in equating. In addition, the use of constructed-response items and computer-based tests became more prominent. These applications create complexities for equating beyond what is typically encountered with paper-and-pencil multiple-choice tests. Thus, updating the material in the first edition was one of the reasons for publishing a second edition.

The first edition briefly considered score scales and test linking. The second edition devoted whole chapters to each of these topics. The development of score scales is an important component of the scaling and equating process. Linking of tests has been of much recent interest, due to various investigations of how to link tests from different test publishers or constructed for different purposes (e.g., Feuer et al. 1999). Because both scaling and linking are closely related to test equating, it seemed natural to extend coverage along these lines.

Following the publication of the second edition in 2004, a considerable amount of research was conducted on equating, scaling, and linking. In addition to a substantial number of journal articles, Dorans, Pommerich, and Holland (2007) and von Davier (2011) published edited books on equating, scaling, and linking. In addition, a substantial chapter by Holland and Dorans (2006) provides a conceptual framework for classifying equating and linking methodology that focuses on the properties of scores that are linked and on the requirements of different types of linking. A chapter by Kolen (2006) provides a updated discussion of score scales. The third edition updates all chapters to incorporate this recent literature. Following is a brief overview of the chapters of the third edition.

In Chap. 1, a general introduction is provided, primarily in terms of a conceptual overview. In this chapter, we define equating, describe its relationship to test development, and distinguish equating from scaling and linking. We also present equating designs, properties of equating, and introduce the concept of equating error.

In Chap. 2, using the random groups design, we illustrate traditional equating methods, such as equipercentile and linear methods. We also discuss here many of the key concepts of equating, such as properties of converted scores and the influence of the resulting scale scores on the choice of an equating result.

In Chap. 3, we cover smoothing methods in equipercentile equating. We show that the purpose of smoothing is the reduction of random error in estimating equating relationships in the population. We describe methods based on log-linear models, cubic splines, and strong true score models.

In Chap. 4, we treat linear equating with nonequivalent groups of examinees. We derive statistical methods and stress the need to disconfound examinee-group and test-form differences. Also, we distinguish observed score equating from true score equating.

In Chap. 5, we continue our discussion of equating with nonequivalent groups with a presentation of equipercentile methods.

In Chap. 6, we describe item response theory (IRT) equating methods under various designs. This chapter covers issues that include scaling person and item parameters, IRT true and observed score equating methods, equating using item pools, and equating using polytomous IRT models.

Chapter 7 focuses on standard errors of equating; both bootstrap and analytic procedures are described. We illustrate the use of standard errors to choose sample sizes for equating and to compare the precision in estimating equating relationships for different designs and methods.

In Chap. 8, we describe many practical issues in equating, including the importance of test development procedures, test standardization conditions, and quality control procedures. We stress conditions that are conducive to adequate equating. Also, we discuss comparability issues for mixed-format assessments and computer-based tests.

Chapter 9 is devoted to score scales for tests. We discuss different scaling perspectives. We describe linear and nonlinear transformations that are used to construct score scales, and we consider procedures for enhancing the meaning of scale scores that include incorporating normative, content, and score precision information. We discuss procedures for maintaining score scales and scales for batteries and composites. We conclude with a section on vertical scaling that includes consideration of scaling designs and psychometric methods and a review of research on vertical scaling.

In Chap. 10, we describe linking categorization schemes and criteria and consider equating, vertical scaling, and other related methodologies as a part of these categorization schemes. An extensive example is used to illustrate how the lack of group invariance in concordance relationships can be examined and used as a means for demonstrating some of the limitations of linking methods.

We use a random groups illustrative equating example in Chaps. 2, 3, and 7; a nonequivalent groups illustrative example in Chaps. 4–6; a second random groups illustrative example in Chaps. 6 and 9; and a single-group illustrative example in Chap. 10. We use data from the administration of a test battery in multiple grades for an illustrative example in Chap. 9, and data from the administration of two different tests for an illustrative example in Chap. 10. Chapters 1–10 each have a set of exercises that are intended to reinforce the concepts and procedures in the chapter. The answers to the exercises are in Appendix A. We describe computer programs and how to obtain them in Appendix B.

In addition to updating the review of literature for all of the chapters, the third edition incorporates substantial new material as follows:

- Chapter 3 includes additional procedures to choose models in log-linear pre-smoothing and includes a new brief section on the kernel method of equating.
- Chapter 4 includes a new section on chained linear equating and incorporates chained linear equating in the illustrative example. In addition, it includes a new

discussion of the relationships among linear methods in the common-item nonequivalent groups design.
- Chapter 5 includes new descriptions of modified frequency estimation equating and chained equipercentile equating, and incorporates these methods in the illustrative example.
- Chapter 8 includes a new extensive section on equating criteria in research studies. Material on equating mixed-format tests containing multiple-choice and constructed-response items is significantly updated.
- Chapter 9 includes a new section on unit scores, item scores, and raw scores. A new section on scores for mixed-format tests, including issues in weighting scores for different item types, is added. In addition, a new section on score scales and growth is added.
- Chapter 10 includes a new summary of the Holland and Dorans (2006) linking framework.

In addition, each chapter contains a reference list, rather than having a single reference list at the end of the volume as in the first two editions.

We anticipate that many readers of this book will be advanced graduate students, entry-level professionals, or persons preparing to conduct equating, scaling, or linking for the first time. Other readers likely will be experienced professionals in measurement and related fields who will want to use this book as a reference. To address these varied audiences, we make frequent use of examples and stress conceptual issues. This book is not a traditional statistics text. Instead, it is meant for instructional use and as a reference for practical use that is intended to address both statistical and applied issues. The most frequently used methodologies are treated, as well as many practical issues. Although we are unable to cover all of the literature on equating, scaling, and linking, we provide many references so that the interested reader may pursue topics of particular interest.

The principal goals of this book are for the reader to understand the principles of equating, scaling, and linking; to be able to conduct equating, scaling, and linking; and to interpret the results in reasonable ways. After studying this book, the reader should be able to

- Understand the purposes of equating, scaling, and linking and the context in which they are conducted.
- Distinguish between equating, scaling, and linking methodologies and procedures.
- Appreciate the importance to equating of test development and quality control procedures.
- Understand the distinctions among equating properties, equating designs, and equating methods.
- Understand fundamental concepts—including designs, methods, errors, and statistical assumptions.
- Compute equating, scaling, and linking functions and choose among methods.
- Interpret results from equating, scaling, and linking analyses.

- Design reasonable and useful equating, scaling, and linking studies.
- Conduct equating, scaling, and linking in realistic testing situations.
- Identify appropriate and inappropriate uses and interpretations of equating, scaling, and linking results.

We cover nearly all of the material in this book in a three semester-hour graduate seminar at The University of Iowa. In our course, we supplement the materials here with general references (Angoff 1971; Holland and Dorans 2006; Holland and Rubin 1982; Petersen et al. 1989) so that the students become familiar with other perspectives and notational schemes.

We have used much of the material in this book in various training sessions, including those at the annual meetings of the National Council on Measurement in Education, the American Educational Research Association, and the American Psychological Association, and in workshops given in Israel, Japan, South Korea, Spain, Taiwan, and The University of Iowa.

We acknowledge the generous contributions that others made to the first edition of this book. We benefitted from interactions with very knowledgeable psychometricians at ACT and elsewhere, and many of the ideas in this book came from conversations and interactions with these people. Specifically, Bradley Hanson reviewed the entire manuscript and made valuable contributions, especially to the statistical presentations. He conducted the bootstrap analyses that are presented in Chapter 7 and, along with Lingjia Zeng, developed much of the computer software used in the examples. Deborah Harris reviewed the entire manuscript, and we thank her especially for her insights on practical issues in equating. Chapters 1 and 8 benefitted considerably from her ideas and counsel. Lingjia Zeng also reviewed the entire manuscript and provided us with many ideas on statistical methodology, particularly in the areas of standard errors and IRT equating. Thanks to Dean Colton for his thorough reading of the entire manuscript, Xiaohong Gao for her review and for working through the exercises, and Ronald Cope and Tianqi Han for reading portions of the manuscript. We are grateful to Nancy Petersen for her in-depth review of a draft of the first edition, her insights, and her encouragement. Bruce Bloxom provided valuable feedback, as did Barbara Plake and her graduate class at the University of Nebraska–Lincoln. We thank an anonymous reviewer, and the reviewer's graduate student, for providing us with their valuable critique. We are indebted to all who have taken our equating courses and training sessions.

For the second edition, we are grateful to Ye Tong for the many hours she spent on electronic typesetting, for all of the errors she found, and for helping with many of the examples and the exercises. We thank Amy Hendrickson for helping to develop the polytomous IRT examples in Chapter 6, Seonghoon Kim for reviewing the additions to Chapter 6 on polytomous IRT and for developing the computer program POLYST, and Ping Yin for her work on Chapters 4 and 10. We acknowledge the work of Zhongmin Cui and Yueh-Mei Chien on the computer programs, and the work of Noo Ree Huh on checking references. We thank the students in our equating and scaling classes at The University of Iowa who discovered many errors and for helping us clarify some confusing portions of earlier drafts. We are grateful to Neil

Dorans, Samuel Livingston, and Paul Holland for reviewing portions of the new material in the second edition. We express our appreciation to the Iowa Measurement Research Foundation for providing support to the graduate students who worked with us on the second edition. For the third edition, we thank Wei Wang for her many hours spent on electronic typesetting. We also thank many graduate students at The University of Iowa for helping us correct errors that appeared in the second edition. Amy Kolen deserves thanks for her superb editorial advice for all three editions.

Iowa City, IA November, 2013

Michael J. Kolen
Robert L. Brennan

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Dorans, N. J., Pommerich, M., & Holland, P. (Eds.). (2007). *Linking and aligning scores and scales*. New York: Springer.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Research Council.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, *6*, 195–240.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: American Council on Education and Praeger.

Holland, P. W., & Rubin, D. B. (1982). *Test equating*. New York: Academic.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–186). Westport, CT: American Council on Education and Praeger.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan.

von Davier, A. A. (Ed.). (2011). *Statistical models for test equating, scaling, and linking*. New York: Springer.

Contents

Notation

Arabic

1	Population taking Form X (Chapter 4)
2	Population taking Form Y (Chapter 4)
A	Slope constant in linear equating and raw-to-scale score transformations (Chapter 4)
A	Slope constant in IRT θ scale transformation (Chapter 6)
a	Item slope parameter in IRT (Chapter 6)
B	Location constant in linear equating and raw-to-scale score transformations (Chapter 4)
B	Location constant in IRT θ scale transformation (Chapter 6)
b	Item location parameter in IRT (Chapter 6)
b	Item or category location parameter in polytomous IRT (Chapter 6)
b^*	Nonlinear transformation of b (Chapter 9)
bias	Bias (Chapter 3)
C	Number of degrees of the polynomial in log-linear smoothing (Chapter 3)
c	Item pseudochance level parameter in IRT (Chapter 6)
c	Item location parameter in Bock's nominal categories model (Chapter 6)
constant	A constant (Chapter 2)
cov	Sampling covariance (Chapter 7)
D	Scaling constant in IRT, usually set to 1.7 (Chapter 6)
DTM	Difference That Matters (Chapter 10)
d	Category location parameter in generalized partial credit model (Chapter 6)
$d_Y(x)$	Expected value of a cubic spline estimator of $e_Y(x)$ (Chapter 3)
$d^*_Y(x)$	Average of two splines (Chapter 3)
df	Degrees of freedom (Chapter 3)
E	Expected value (Chapter 1)
E	Number correct error score (Chapter 4)
e	The equipercentile equating function, such as $e_Y(x)$ (Chapter 2)

$e_Y(x)$	The Form Y equipercentile equivalent of a Form X score (Chapter 1)
$e_X(y)$	The Form X equipercentile equivalent of a Form Y score (Chapter 2)
effect size	Effect size (Chapter 9)
eq	General equating function, such as $eq_Y(x)$ (Chapter 1)
ew	Effective weight (Chapter 9)
ewMAD	Equally weighted average of absolute differences (Chapter 10)
ewMD	Equally weighted average of differences (Chapter 10)
ewREMSD	Equally weighted Root Expected Mean Square Difference (Chapter 10)
exp	Exponential (Chapter 6)
$F(x)$	$Pr(X \leq x)$ is the cumulative distribution for X (Chapter 1)
F^*	Cumulative distribution function of $eq_X(y)$ (Chapter 2)
F^{-1}	Inverse of function F (Chapter 2)
f	A general function (Chapter 7)
f'	The first derivative of f (Chapter 7)
$f(x)$	$Pr(X = x)$ is the discrete density for X (Chapter 2)
$f(x,v)$	$Pr(X = x \, and \, V = v)$ is the joint density of X and V (Chapter 5)
$f(x\|v)$	$Pr(X = x \, given \, V = v)$ is the conditional density of x given v (Chapter 5)
func	Function solved for in Newton–Raphson iterations (Chapter 6)
func′	First derivative of function solved for in Newton–Raphson iterations (Chapter 6)
$G(y)$	$Pr(Y \leq y)$ is the cumulative distribution for Y (Chapter 1)
G^*	The cumulative distribution function of e_Y (Chapter 1)
G^{-1}	Inverse of function G (Chapter 2)
g	Item subscript in IRT (Chapter 6)
g	Index used to sum over categories in generalized partial credit model (Chapter 6)
g	Arcsine transformed proportion-correct score (Chapter 9)
$g(y)$	$Pr(Y = y)$ is the discrete density for Y (Chapter 2)
$g(y,v)$	$Pr(Y = y \, and \, V = v)$ is the joint density of Y and V (Chapter 5)
$g(y\|v)$	$Pr(Y = y \, given \, V = v)$ is the conditional density of y given v (Chapter 5)
g_{adj}	Density adjusted by adding 10^{-6} to each density and then standardizing (Chapter 2)
H	Number of subgroups (Chapter 10)
Hcrit	Criterion function for Haebara's method (Chapter 6)
Hdiff	Difference function for Haebara's method (Chapter 6)
h	Index for summing over categories (Chapter 6)
h	Number of scale score points for a confidence interval (Chapter 9)
h	Subgroup designator (Chapter 10)
$h(v)$	$Pr(V = v)$ is the discrete density for V (Chapter 5)
I	IRT scale (Chapter 6)
I	Number of scale scores on Test X (Chapter 10)
i and i'	Individuals (Chapter 6)

intercept	Intercept of an equating function (Chapter 2)
irt	IRT true-score equating function (Chapter 6)
J	IRT scale (Chapter 6)
J	Number of scale scores on Test Y (Chapter 10)
j and j'	Items (Chapter 6)
K	Number of items (Chapter 2)
KR-20	Kuder–Richardson Formula 20 reliability coefficient (Chapter 9)
KR-21	Kuder–Richardson Formula 21 reliability coefficient (Chapter 9)
k	Lord's k in the Beta4 method (Chapter 3)
k	Categories for an item in polytomous IRT (Chapter 6)
ku	Kurtosis, such as $ku(X) = E[X - \mu(X)]^4\sigma^4(X)$ (Chapter 2)
$l_Y(x)$	The Form Y linear equivalent of a Form X score (Chapter 2)
$l_X(y)$	The Form X linear equivalent of a Form Y score (Chapter 2)
MAD	Weighted average of absolute differences (Chapter 10)
MD	Weighted average of differences (Chapter 10)
m	Number of categories for an item in polytomous IRT (Chapter 6)
$m_Y(x)$	The mean equating equivalent of a Form X score (Chapter 2)
$m_X(y)$	The mean equating equivalent of a Form Y score (Chapter 2)
max	Maximum score (Chapter 6)
min	Minimum score (Chapter 6)
mse	Mean squared error (Chapter 3)
N	Number of examinees (Chapter 2)
NCE	Normal Curve Equivalent (unrounded) (Chapter 9)
NCE_{int}	Normal Curve Equivalent rounded to an integer (Chapter 9)
$P(x)$	The percentile rank function for X (Chapter 2)
$P*$	A given percentile rank (Chapter 2)
$P**$	$P/100$ (Chapter 7)
P^{-1}	The percentile function for X (Chapter 2)
p	Probability of a correct response in IRT (Chapter 6)
p	Category response function in polytomous IRT (Chapter 6)
$p*$	Cumulative category response function in polytomous IRT (Chapter 6)
p'	First derivative of p (Chapter 6)
pl_{Yh}	Parallel linear equating equivalent on Test Y for subgroup h (Chapter 10)
$Q(y)$	Percentile rank function for Y (Chapter 2)
Q^{-1}	Percentile function for Y (Chapter 2)
R	Number of bootstrap replications (Chapter 7)
REMSD	Root Expected Mean Square Difference (Chapter 10)
RMSD	Root Mean Square Difference (Chapter 10)
RP	Response Probability level in item mapping (Chapter 9)
r	Index for calculating observed score distribution in IRT (Chapter 6)
r	Index for bootstrap replications (Chapter 7)
rmsel	Root mean squared error for linking (Chapter 10)

S	Smoothing parameter in postsmoothing (Chapter 3)
SC	Scale score random variable (Chapter 9)
$SLcrit$	Criterion function for Stocking-Lord method (Chapter 6)
$SLdiff$	Difference function for Stocking-Lord method (Chapter 6)
SMD	Standardized Mean Difference (Chapter 10)
s	Synthetic population (Chapter 4)
sc	Scale score transformation, such as $sc(y)$ (Chapter 2)
sc_{int}	Scale score rounded to an integer (Chapter 2)
se	Standard error, such as $se(x)$ (Chapter 3)
sem	Standard error of measurement (Chapter 7)
sk	Skewness, such as $sk(X) = E[X - \mu(X)]^3/\sigma^3(X)$ (Chapter 2)
$slope$	Slope of equating function (Chapter 2)
st	Stanine (unrounded) (Chapter 9)
st	Scaling test (Chapter 9)
st_{int}	Stanine rounded to an integer (Chapter 9)
T	Number correct true score (Chapter 4)
T	Normalized score with mean of 50 and standard deviation of 10 (Chapter 9)
T_{int}	Normalized score with mean of 50 and standard deviation of 10 rounded to an integer (Chapter 9)
t	Realization of number correct true score (Chapter 4)
$t_Y(x)$	Expected value of an alternate estimator of $e_Y(x)$ (Chapter 3)
U	Uniform random variable (Chapter 2)
u	Standard deviation units (Chapter 7)
V	The random variable indicating raw score on Form V (Chapter 4)
v	Spline coefficient (Chapter 3)
v	A realization of V (Chapter 4)
v	Subgroup weight for a particular score (Chapter 10)
var	Sampling variance (Chapter 3)
w	Weight for synthetic group (Chapter 4)
w	Nominal weight (Chapter 9)
w	Subgroup weight (Chapter 10)
X	The random variable indicating raw score on Form X (Chapter 1)
X	Random variable indicating scale score on Test X (Chapter 10)
$X*$	Equals $X+ U$, used in the continuization process (Chapter 2)
x	A realization of X (Chapter 2)
$x*$	Integer closest to x such that $x^* - .5 \leq x < x^* + .5$ (Chapter 2)
$x*$	Form X_2 score equated to the Form X_1 scale (Chapter 7)
x_{high}	Upper limit in spline calculations (Chapter 3)
x^*_L	The largest integer score with a cumulative percent less than $P*$ (Chapter 2)
x_{low}	Lower limit in spline calculations (Chapter 3)
x^*_U	Smallest integer score with a cumulative percent greater than $P*$ (Chapter 2)

Y	The random variable indicating raw score on Form Y (Chapter 1)
Y	Random variable indicating scale score on Test Y (Chapter 10)
y	A realization of Y (Chapter 1)
y_i^*	Largest tabled raw score less than or equal to $e_Y(x)$ in finding scale scores (Chapter 2)
y_L^*	The largest integer score with a cumulative percent less than Q^* (Chapter 2)
y_U^*	The smallest integer score with a cumulative percent greater than Q^* (Chapter 2)
Z	The random variable indicating raw score on Form Z (Chapter 4)
z	A realization of Z (Chapter 4)
z	Unit normal variable (Chapter 7)
z	Normalized score (Chapter 10)
z^*	Selected set of normalized scores in Thurstone scaling (Chapter 9)
z_γ	Unit normal score associated with a 100γ % confidence interval (Chapter 9)

Greek

$\alpha(X\|V)$	Linear regression slope (Chapter 4)
$\alpha(Y\|V)$	Linear regression slope (Chapter 4)
$\beta(X\|V)$	Linear regression intercept (Chapter 4)
$\beta(Y\|V)$	Linear regression intercept (Chapter 4)
χ^2	Chi-square test statistic (Chapter 3)
δ	Location parameter in congeneric models (Chapter 4)
ϕ	Normal ordinate (Chapter 7)
γ	Expansion factor in linear equating with the common-item nonequivalent groups design (Chapter 4)
γ	Confidence coefficient (Chapter 9)
λ	Effective test length in congeneric models (Chapter 4)
μ	Mean as in $\mu(X)$ and $\mu(Y)$ (Chapter 2)
ν	Weight for a pair of subgroups and a particular score (Chapter 10)
Φ	Inverse normal transformation (Chapter 9)
Θ	Parameter used in developing the delta method (Chapter 7)
θ	Ability in IRT (Chapter 6)
θ^+	New value in Newton–Raphson iterations (Chapter 6)
θ_-	Initial value in Newton–Raphson iterations (Chapter 6)
θ^*	Nonlinear transformation of θ (Chapter 9)
ρ	Correlation, such as $\rho(X, V)$ (Chapter 4)
$\rho(X, X')$	Reliability (Chapter 4)
$\sigma(X, V)$	Covariance between X and V (Chapter 4)
$\sigma(Y, V)$	Covariance between Y and V (Chapter 4)
σ^2	Variance such as $\sigma^2(X) = \mathbf{E}[X - \mu(X)]^2$ (Chapter 4)
σ_{ij}	Covariance between variables i and j (Chapter 9)

τ	True score (Chapter 1)
τ^*	True score outside of range of possible true scores (Chapter 6)
$\hat{\tau}$	Estimated true scores (Chapter 9)
ω	Weight in log-linear smoothing (Chapter 3)
Ψ	Function that relates true scores (Chapter 4)
ψ	Distribution of a latent variable (Chapter 3)
∂	Partial derivative (Chapter 7)

Chapter 1
Introduction and Concepts

This chapter provides a general overview of equating and briefly considers important concepts. The concept of equating is described, as is why it is needed, and how to distinguish it from other related processes. Equating properties and designs are considered in detail, because these concepts provide the organizing themes for addressing the statistical methods treated in subsequent chapters. Some issues in evaluating equating are also considered. The chapter concludes with a preview of subsequent chapters.

1.1 Equating and Related Concepts

Scores on tests often are used as one piece of information in making important decisions. Some of these decisions focus at the *individual level*, such as when a student decides which college to attend or on a course in which to enroll. For other decisions the focus is more at an *institutional level*. For example, an agency or institution might need to decide what test score is required to certify individuals for a profession or to admit students into a college, university, or the military. Still other decisions are made at the *public policy level*, such as addressing what can be done to improve education in the United States and how changes in educational practice can be evaluated. Regardless of the type of decision that is to be made, it should be based on the most accurate information possible: All other things being equal, *the more accurate the information, the better the decision*.

Making decisions in many of these contexts requires that tests be administered on multiple occasions. For example, college admissions tests typically are administered on particular days, referred to as *test dates*, so examinees can have some flexibility in choosing when to be tested. Tests also are given over many years to track educational trends over time. If the same test questions were routinely administered on each test

Some of the material in this chapter is based on Kolen (1988).

M. J. Kolen and R. L. Brennan, *Test Equating, Scaling, and Linking*,
Statistics for Social and Behavioral Sciences, DOI: 10.1007/978-1-4939-0317-7_1,

date, then examinees might inform others about the test questions. Or, an examinee who tested twice might be administered the same test questions on the two test dates. In these situations, a test might become more of a measure of exposure to the specific questions that are on the test than of the construct that the test is supposed to measure.

1.1.1 Test Forms and Test Specifications

These test security problems can be addressed by administering a different collection of test questions, referred to as a *test form*, to examinees who test on different test dates. A test form is a set of test questions that is built according to content and statistical *test specifications* (Schmeiser and Welch 2006). Test specifications provide guidelines for developing the test. Those responsible for constructing the test, the *test developers*, use these specifications to ensure that the test forms are as similar as possible to one another in content and statistical characteristics.

1.1.2 Equating

The use of different test forms on different test dates leads to another concern: the forms might differ somewhat in difficulty. *Equating* is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably. Equating adjusts for differences in difficulty among forms that are built to be similar in difficulty and content.

The following situation is intended to develop further the concept of equating. Suppose that a student takes a college admissions test for the second time and earns a higher reported score than on the first testing. One explanation of this difference is that the reported score on the second testing reflects a higher level of achievement than the reported score on the first testing. However, suppose that the student had been administered exactly the same test questions on both testings. Rather than indicating a higher level of achievement, the student's reported score on the second testing might be inflated because the student had already been exposed to the test items. Fortunately, a new test form is used each time a test is administered for most college admissions tests. Therefore, a student would not likely be administered the same test questions on any two test dates.

The use of different test forms on different test dates might cause another problem, as is illustrated by the following situation. Two students apply for the same college scholarship that is based partly on test scores. The two students take the test on different test dates, and Student 1 earns a higher reported score than Student 2. One possible explanation of this difference is that Student 1 is higher achieving than Student 2. However, if Student 1 took an easier test form than Student 2, then Student 1 would have an unfair advantage over Student 2. In this case, the difference in scores might be due to differences in the difficulty of the test forms rather than in

the achievement levels of the students. To avoid this problem, equating is used with most college admissions tests. If the test forms are successfully equated, then the difference in equated scores for Student 1 and Student 2 is not attributable to Student 1's taking an easier form.

The process of equating is used in situations where such *alternate forms* of a test exist and scores earned on different forms are compared to each other. Even though test developers attempt to construct test forms that are as similar as possible to one another in content and statistical specifications, the forms typically differ somewhat in difficulty. Equating is intended to adjust for these difficulty differences, allowing the forms to be used interchangeably. *Equating adjusts for differences in difficulty, not for differences in content*. After successful equating, for example, examinees who earn an equated score of, say, 26 on one test form could be considered, on average, to be at the same achievement level as examinees who earn an equated score of 26 on a different test form.

1.1.3 Processes That are Related to Equating

There are processes that are similar to equating, which may be more properly referred to as *scaling to achieve comparability*, in the terminology of the *Standards for Educational and Psychological Testing* (AERA, APA, NCME 1999), or *linking*, in the terminology of Holland and Dorans (2006), Linn (1993) and Mislevy (1992). One of these processes is *vertical scaling* (frequently referred to as *vertical "equating"*), which often is used with elementary school achievement test batteries. In these batteries, students often are administered questions that test content matched to their current grade level. This procedure allows developmental scores (e.g., grade equivalents) of examinees at different grade levels to be compared. Because the content of the tests administered to students at various educational levels is different, however, scores on tests intended for different educational levels cannot be used interchangeably. Other examples of linking include relating scores on one test to those on another, and scaling the tests within a battery so that they all have the same distributional characteristics. As with vertical scaling, solutions to these problems do not allow test scores to be used interchangeably, because the content of the tests is different.

Although similar statistical procedures often are used in linking and equating, their purposes are different. Whereas tests that are purposefully built to be different are linked, equating is used to adjust scores on test forms that are built to be as similar as possible in content and statistical characteristics. When equating is successful, scores on alternate forms can be used interchangeably. Issues in linking tests that are not built to the same specifications are considered further in Chaps. 9 and 10.

1.1.4 Equating and Score Scales

On a multiple-choice test, the *raw score* an examinee earns is often the number of items the examinee answers correctly. Other raw scores might involve penalties for wrong answers or weighting items differentially. On tests that require ratings by judges, a raw score might be the sum of the numerical ratings made by the judges.

Raw scores often are transformed to *scale scores*. The *raw-to-scale score transformation* can be chosen by test developers to enhance the interpretability of scores by incorporating useful information into the score scale (Kolen 2006; Petersen et al. 1989). Information based on a nationally representative group of examinees, referred to as a *national norm group*, sometimes is used as a basis for establishing score scales. For example, the number-correct scores for the four tests of the initial form of a revised version of the ACT tests were scaled (Brennan 1989) to have a mean scale score of 18 for a nationally representative sample of college-bound 12th graders. Thus, an examinee who earned a scale score of 22, for example, would know that this score was above the mean scale score for the nationally representative sample of college-bound 12th graders used to develop the score scale. One alternative to using nationally representative norm groups is to base scale score characteristics on a *user norm group*, which is a group of examinees that is administered the test under operational conditions. For example, a rescaled SAT scale was established for use beginning in 1995 by setting the mean score equal to 500 for the group of SAT examinees that graduated from high school in 1990 (Cook 1994; Dorans 2002).

Scaling and Equating Process

Equating can be viewed as an aspect of a more general *scaling and equating process*. Score scales typically are established using a single test form. For subsequent test forms, the scale is maintained through an equating process that places raw scores from subsequent forms on the established score scale. In this way, a scale score has the same meaning regardless of the test form administered or the group of examinees tested. Typically, raw scores on the new form are equated to raw scores on the old form, and these equated raw scores are then converted to scale scores using the raw-to-scale score transformation for the old form.

Example of the Scaling and Equating Process

The hypothetical conversions shown in Table 1.1 illustrate the scaling and equating process. The first two columns show the relationship between Form Y raw scores and scale scores. For example, a raw score of 28 on Form Y converts to a scale score of 14 (At this point there is no need to be concerned about what particular method was used to develop the raw-to-scale score transformation). The relationship between Form

Table 1.1 Hypothetical conversion tables for test forms

Scale	Form Y raw	Form X_1 raw	Form X_2 raw
•	•	•	•
•	•	•	•
•	•	•	•
13	26	27	28
14	27	28	29
14	28	29	30
15	29	30	31
15	30	31	32
•	•	•	•
•	•	•	•
•	•	•	•

Y raw scores and scale scores shown in the first two columns involves scaling—not equating, because Form Y is the only form that is being considered so far.

Assume that an equating process indicates that Form X_1 is 1 point easier than Form Y throughout the score scale. A raw score of 29 on Form X_1 would thus reflect the same level of achievement as a raw score of 28 on Form Y. This relationship between Form Y raw scores and Form X_1 raw scores is displayed in the second and third columns in Table 1.1. What scale score corresponds to a Form X_1 raw score of 29? A scale score of 14 corresponds to this raw score, because a Form X_1 raw score of 29 corresponds to a Form Y raw score of 28, and a Form Y raw score of 28 corresponds to a scale score of 14.

To carry the example one step further, assume that Form X_2 is found to be uniformly 1 raw score point easier than Form X_1. Then, as illustrated in Table 1.1, a raw score of 30 on Form X_2 corresponds to a raw score of 29 on Form X_1, which corresponds to a raw score of 28 on Form Y, which corresponds to a scale score of 14. Later, additional forms could be converted to scale scores by a similar chaining process. The result of a successful scaling and equating process is that scale scores on all forms can be used interchangeably.

Possible Alternatives to Equating

Equating has the potential to improve score reporting and interpretation of tests that have alternate forms when examinees administered different forms are evaluated at the same time, or when score trends are to be evaluated over time. When at least one of these characteristics is present, at least two possible, but typically unacceptable, alternatives to equating exist. One alternative is to report raw scores regardless of the form administered. As was the case with Students 1 and 2 considered earlier, this approach could cause problems because examinees who were administered an easier form are advantaged and those who were administered a more difficult